Data Mining & Warehousing

Data Mining :-

→ Knowledge Mining.

→ Knowledge discovery from Data (KDD)

Extracting the knowledge from the large amount of Data.

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern evaluation
- Knowledge presentation

- Temporal Databases
- Sequence Database
- Time-series Database

Temporal Databases

→ time related attributes.

→ timestamps

Sequence Database :-

→ sequence of events or activity happens

→

Timeseries Database :-

→ sequence of values or events over repeated measurement of time (minute / second / milli second / daily / weekly etc

Ex :-

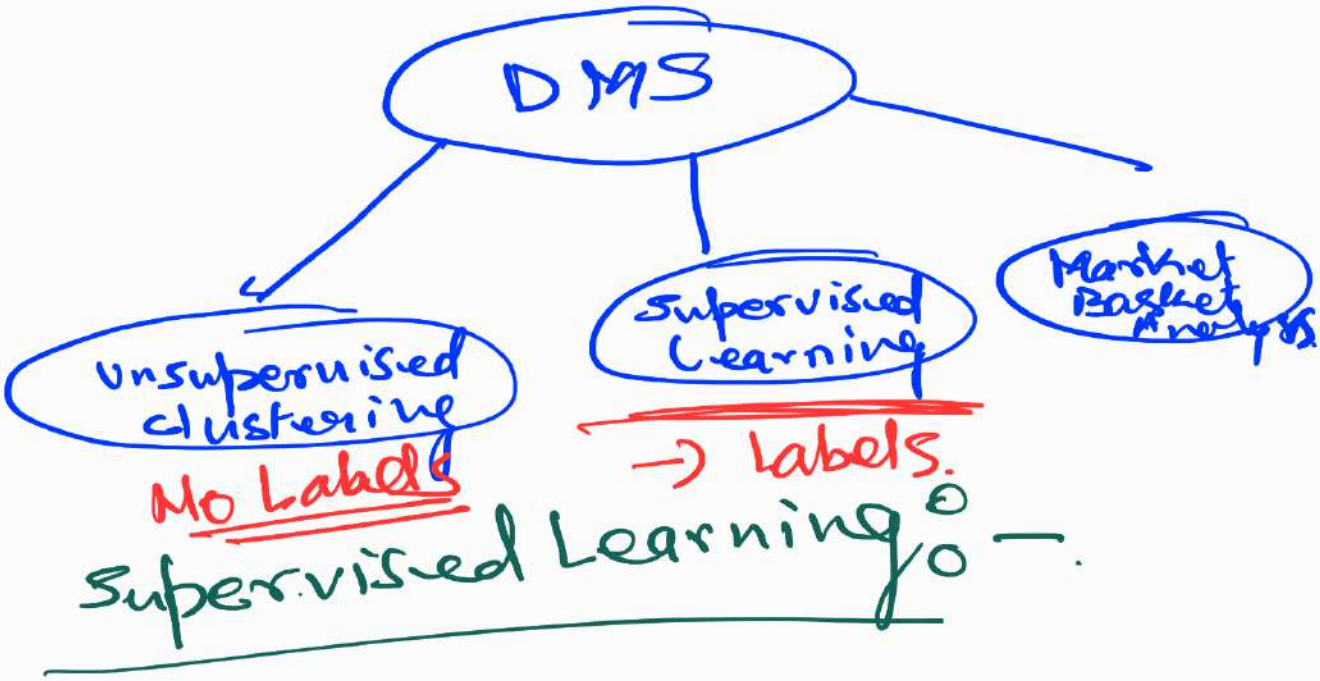
Data Mining

→ Data mining (KDD)
is the process of analyzing data from different perspective & summarizing it into useful information that can be used to
→ increase revenue.
→ cut ~~cost~~ cost

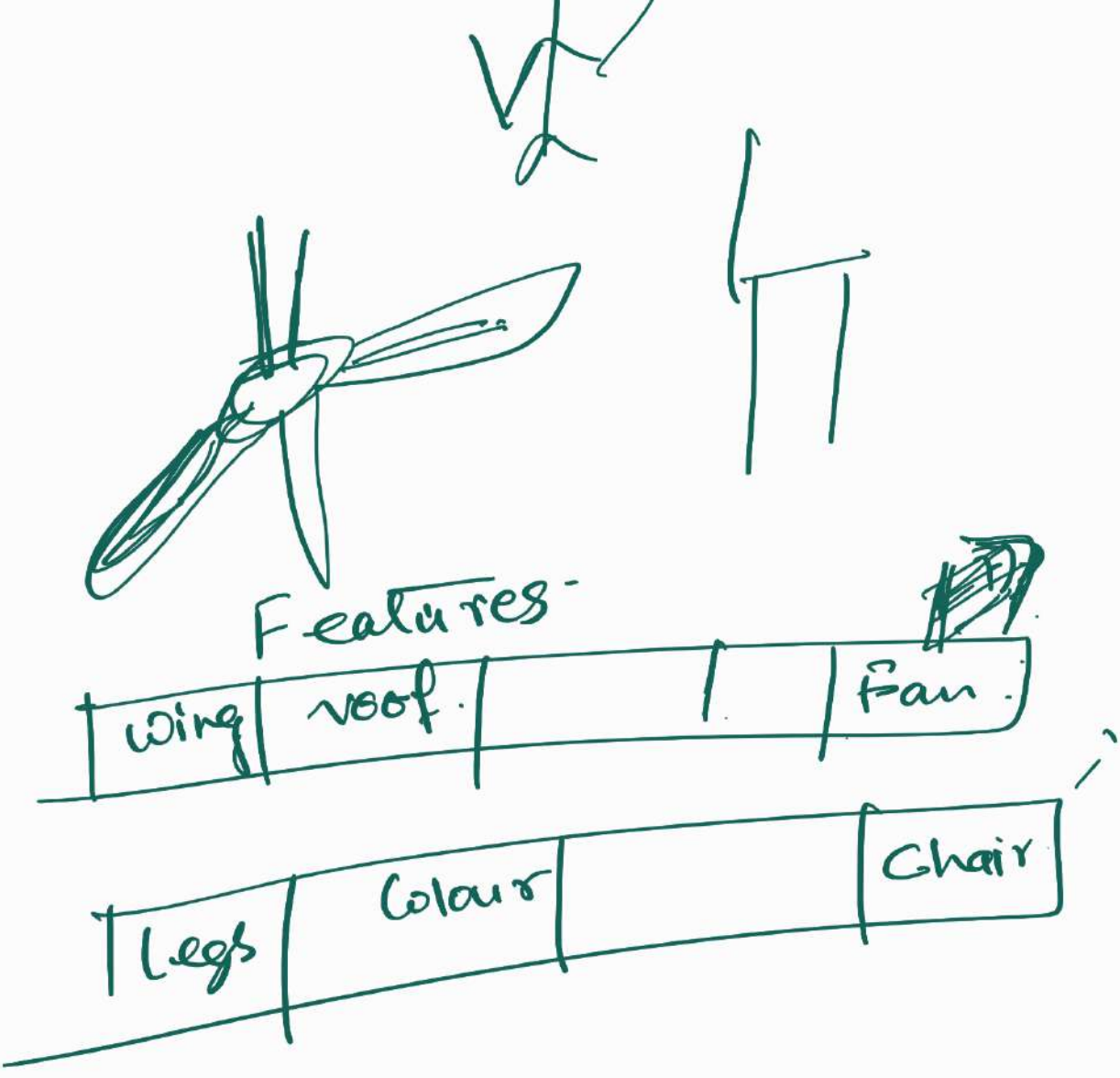
or

Both -

Data Mining strategies



Supervised Learning -



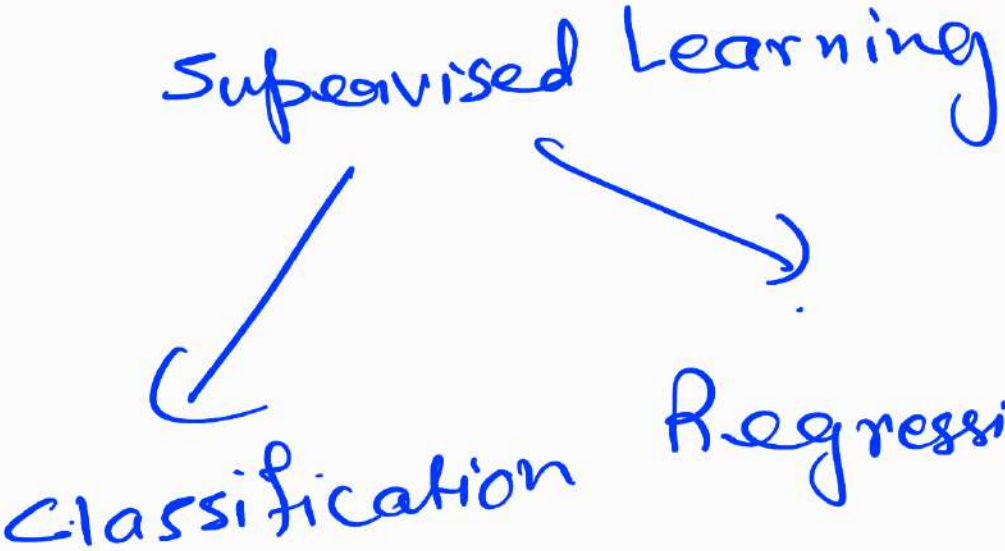
→ When we are young
→ we use induction
to form basic
concept.
→ animals
plants
buildings

→ mind choose what

we believe to be the defining concept feature & form our own classification model.

→ Supervision
  ↳ Labels
  ↳ instances

→ induction based learning.



Classification —

↳ the output is

having certain defined labels (discrete values)

|→ Binary or Multiclass

o/p → Yes | No

Regression :- |→ Red, Green Blue.

|→ here the output is having continuous values

Ex:-

Temp	Press	Humidity	wind speed
10.34	924.4	———	?

→ Predict the sale.

→ Price of house " " share

Unsupervised Clustering :-

|→ it builds model from data without

predefined classes

→ Data instances are grouped together based on the similarity scheme defined by clustering system.

→ Clustering

→ grouping the data in to clusters, so that objects within a cluster have high similarity

→ but are very dissimilar to objects in other clusters.

→ Dissimilarity is accessed based on the attribute values that are describing the object

→ Distance measure is generally used.

→ supervised.

→ Unsupervised.

→ Semi Supervised Learing..

First we apply clustering algo

↳ Labelled it.

↓↓

Labelled Data.

→ Data ~~Mining~~ Mining

↳ steps.

→ Types of DM Alg.

→ Market Basket Analysis :-

Mobile → Mobile Cover
Mobile → Tempered Glass

→ Rules
→ Sales
→ Association Rule

A item → B item

Mobile → Tempered Glass

↓

Arrange.

Mobile, Tempered Glass

10:30 — hrs.

Data Mining

Mean →

Take few numbers

13, 18, 13, 14, 13, 16, 14, 21, 13

mean = ?

$$\frac{\text{Sum of num}}{\text{Total No.}} = \frac{135}{9}$$

Median →

13, 13, 13, 13, 14, 14, 16, 18, 21

find the median ?

Ans ⇒ 14

Median ⇒ it is the middle value of the data set when

it has been arranged in order.

Ex 2 :-

12, 18, 16, 21, 10, 13, 17, 19

if the data set is even.

⇒ median is the average of two middle values

→ 10, 12, 13, 16, 17, 18, 19, 21

$$\Rightarrow \frac{4^{th}\text{data val} + 5^{th}\text{data val}}{2}$$

$$\Rightarrow \frac{16+17}{2} = 16.5$$

Mode :-

48, 44, 48, 45, 42, 49, 48

⇒ data values in the data set that occurs

most often.

Ex$^{n}_{o}$ →

dataset is given like this

age	frequency
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

L_1 → 20-50 ← 1597

Calculate the Approx. median value of the data.

$$\text{median} = L_1 + \left(\frac{N/2 - (\Sigma \text{freq})_1}{\text{freq}_{\text{median}}} \right) \text{width}$$

L_1 = lower boundary of

L_1 ⇒ Lower Boundary of the median interval

N ⇒ is the no. of values in the data set

$(\Sigma freq)_l$ ⇒ sum of frequencies of all the intervals that are lower than the median interval

$freq_{median}$ ⇒ is the freq of the median interval

width ⇒ It is the width of the median interval

32.94	33.4138	34.94
30.94	29.10	

$N = 3194$ median ⇒ 1597

$L_1 = 20$

$(\Sigma freq)_l = 950$

$\Sigma freq_i =$

$freq_{median} = 1500$

$width = 30$

$median = 32.94$

Data Preprocessing

→ Data Cleaning

|→ incomplete,
noisy, inconsistent

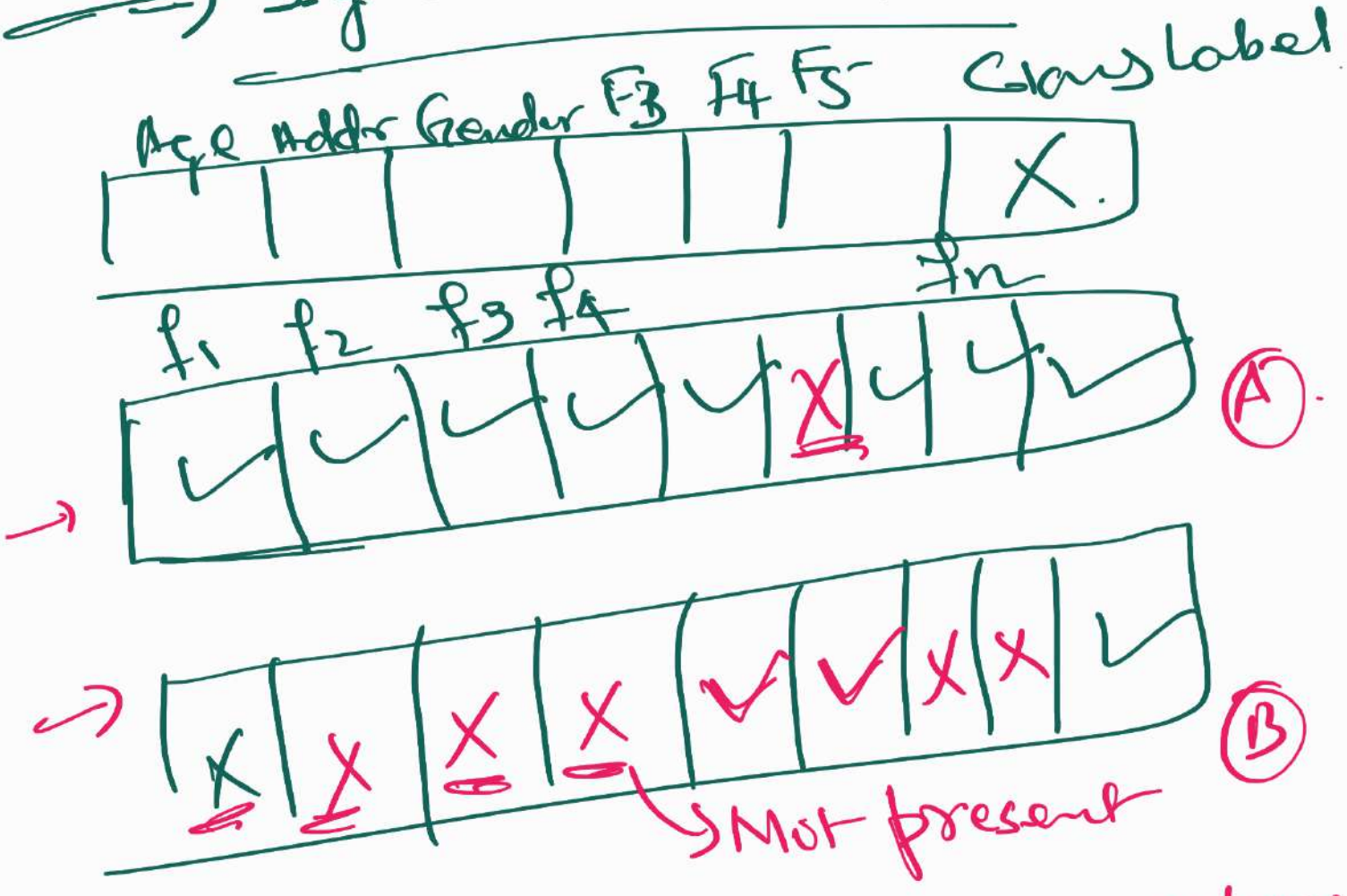
→ fill in the missing values

→ smooth out the noise
while identifying
outliers

→ Correct Inconsistency
in the data

Missing Values

→ Ignore the tuple



→ Fill in the missing values manually :-

→ Use a global constant to fill in the missing value

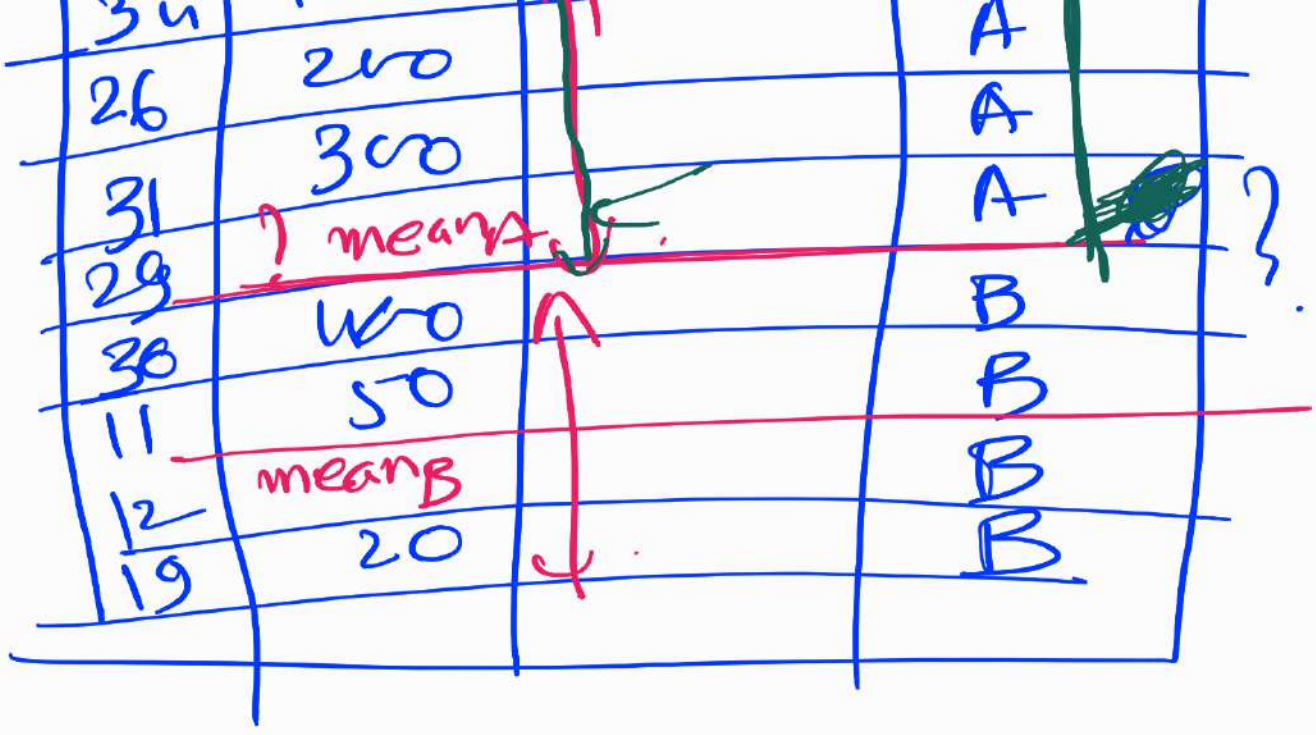
Ex- unknown, 0, ∞ or ∞ etc

- Use the attribute mean to fill in the missing value.

F1	Income		Class
34	100		
26	200		
33	Mean		
28	300		
33	150		
11	200		

- Use the attribute mean for all samples belonging to the same class as the given tuple.

F1	Income	fn	Class
	100		A



⇒ Use the most probable values to fill in the missing value.

→ Eg:- Regression
Bayesian
Decision Tree

Standard Deviation

1, 2, 3, 4, 5, 6

S.D. $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

mean = 3.5

SD = 1.7

→ Data Cleaning

- → Missing values
- → Noisy Data

Noisy Data →

↳ random error

- Binning
- Regression
- Clustering

Binning -

↳ It smooths data values by consulting its neighbourhood. i.e the value around it.

Eg:- 4, 8, 15, 21, 21, 24, 25, 28, 34

→ Put the data values into Buckets | Bins.

→ Local smoothing

↳ smoothing by bin

means

→ smoothing by bin boundaries

Bin1 : 4, 8, 15
Bin2 : 21, 21, 24
Bin3 : 25, 28, 34

Smoothing by bin means

Bin1 : 9, 9, 9
Bin2 : 22, 22, 22
Bin3 : 29, 29, 29

Smoothing by bin boundaries.

Bin1 : 4, 4, 15 min = 4 max = 15
Bin2 : 21, 21, 24 min = 21 max = 24
Bin3 : 25, 25, 34 min = 25 max = 34

& Max values are

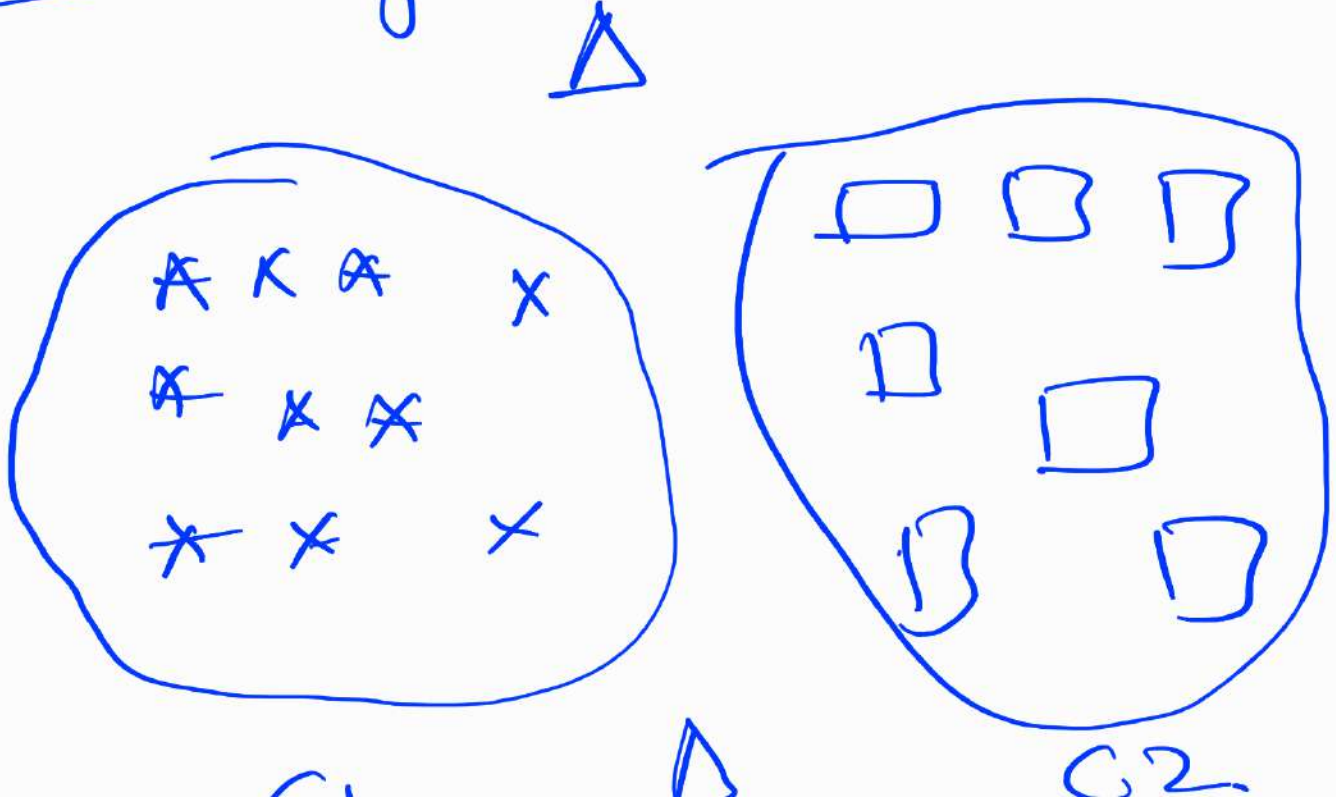
identified ⇒ bin boundaries

→ Each bin values is then replaced by the closest boundary value

Regression –

→ data can be smoothed by fitting it into a function

Clustering –



Outliers

↳ They fall outside the set of clusters

Data Cleaning ..

- → discrepancy detection
- → metadata
- → Data scrubbing tools

↓

use domain knowledge to detect & correct data.

⇒ Data Auditing tool.

⇓

It finds rules & relationship

⇓

Data mining Techniques

Correlation Analysis :-

⇓

relationship b/w Numerical Attributes

⇓

Pearson correlation.

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{A})^2 \sum_{i=1}^{N}(b_i - \bar{B})^2}}$$

$\sqrt{\sum_{i=1}}$

$N \Rightarrow$ No. of tuples

a_i & b_i are respective values of the tuple.

$$-1 \leq r_{A,B} \leq +1.$$

-ve value $\Rightarrow$ Negatively correlated

+ve value $\Rightarrow$ Positively correlated

0 (zero) $\Rightarrow$ A & B are independent & No correlation

Tree Height	Trunk Diameter
4	x

35	8
49	9
27	7
33	6
60	13
21	7
45	11
51	12

Find the correlation b/w Tree Height & Trunk Diameter.

Ans:- $\simeq 0.89$

+ positively correlated

Data Transformation

⊙ ⇒ transformed or consolidated in to

the forms
appropriate for
mining.

1 → smoothing.
→ removal of
noise
→ Binning
→ Regression
→ Clustering

2 → Aggregation -
→ Aggregation o/p.
→ daily → weekly

3 → Generalization
→ we are replacing

the low level data by the higher level concept through the concept hierarchy

Ex:- city can → state

4. → Normalization :-

→ Data are scaled so as to fall with in a specified range

-1 to +1

OR

0 to 1.

5. → Attribute Construction

↳ new attributes

are constructed from
the given set of
attributes.

—— * ——.

Data Transformation

- → Smoothing
- → Aggregation
- → Generalization
- → Normalization.
- → Attribute Construction.

Normalization Techniques

→ Min-max normalization

min_A → min value of an attribute A.

max_A → max. value of an attribute A.

↓

In the new range.

new_min_A → New min val of an attribute A.

new_max$_A$ → new max value of an attribute A.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Ex:- min & max value for attribute income is ₹ 12000/- & ₹ 98000/- map into range. [0.0 to 1.0] & find the value of ₹ 73600/-

0.71 / 0.716

- z-score normalization :-

Zero-mean or mean normalization

value of an attribute A are normalized

↓

based on mean & standard deviation of A.

initial value → V.

After normalization → V'

$$V' = \frac{V - \bar{A}}{\sigma_A}$$

$\bar{A}$ → mean

σ_A → standard deviation

Ex:

attribute income

Mean ⇒ ₹ 154.20/-

SD ⇒ ₹ 16020/-

with z-score non^{n} find the value of ₹ 73600/-

Ans 1.225

Boxplot :-

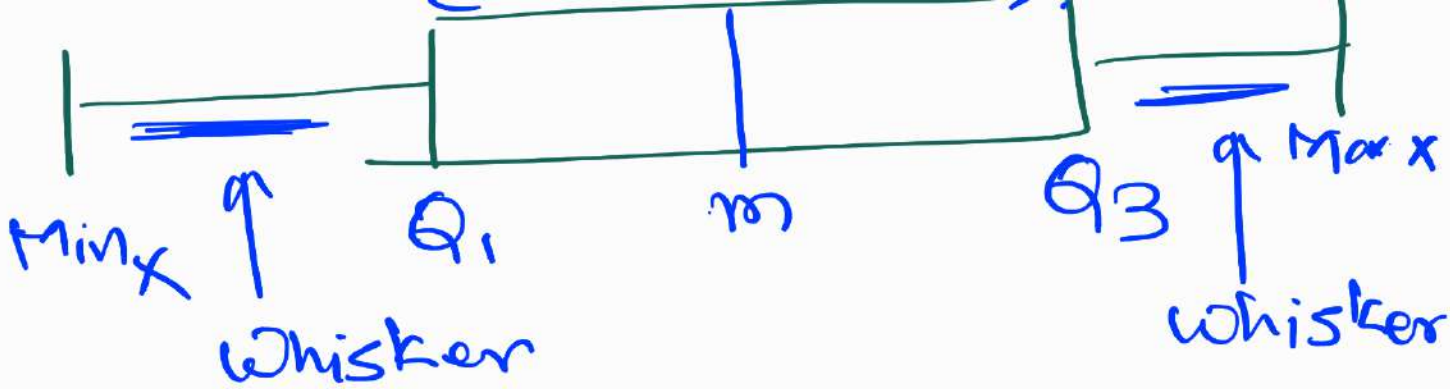
76, 79, 76, 74, 75, 71, 85, 82, 82, 79, 81

→ Arrange the values in the ascending order.

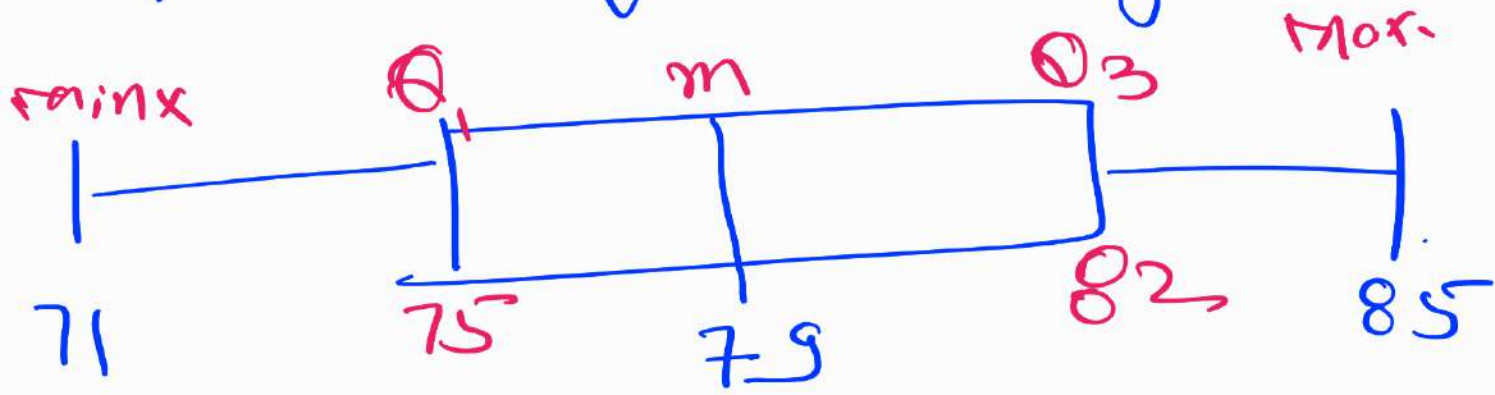
71, 74, 75, 76, 76, 79, 79, 81, 82, 82, 85

→ It gives 5 number summary of the data distribution.

IQR



IQR: Interquartile Range.



71, 74, (75) 76, 76

79, 81, (82) 82, 85

Pearson correlation -

$$-1 \leq r_{AB} \leq 1$$

$-1 \Rightarrow$

$+1 \Rightarrow$

$0 \Rightarrow$

χ^2 (Chi Square Test)

→ It finds the correlation between two categorical attributes

Lets say those attributes are A & B.

→ A ⇒ 'c' distinct values

$a_1, a_2, a_3 \ldots\ldots a_c$

→ B ⇒ 'r' distinct values

$b_1, b_2, b_3 \ldots\ldots b_r$

represented in the form of

'Contingency table'

'c' values of 'A' making up the coln.

&

'r' value of 'B' making up the rows.

(A_i, B_J) → denote an event that

$A \rightarrow a_i$

$B \Rightarrow b_J$

$A = a_i$

$B = b_J$

→ In contingency table we represent different joint event in diff cells

$$\chi^2 \text{ value} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

where O_{ij} is the observed frequency (actual count) of the joint-event (A_i, B_j)

e_{ij} is the expected frequency of (A_i, B_j)

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

$N \rightarrow$ No. of data tuples

$\text{count}(A = a_i) \rightarrow$ No. of tuples having value a_i for A.

$\rightarrow b_j$ for B

Ex:-

	male	female	Total
observed frequency			

fiction	250	(90)	200	(360)	450
non-fiction	50	(210)	1000	(840)	1050
	300.		1200.		1500

expected freq

$$C_{11} = \frac{\text{count(male)} \times \text{count(fiction)}}{N}$$

(male, fiction)

$$= \frac{300 \times 450}{1500} = 90$$

C (fiction, female)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} +$$

$$\frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 507$$

≈ 508

$H_0 \rightarrow$ There is no relationship b/w categorical variable A & categorical variable B.

$H_1 \rightarrow$ There is some relationship b/w these two categorical variables.

Step (2)

Find the degree of freedom

$$(r-1) \times (c-1)$$

$$dof = (2-1) \times (2-1)$$

$$= 1$$

Step (3)

significance level.

1% or 5%

χ^2 , dof , SL

value , — , —

→ If the value in the chart is less than calculated χ^2 then we accept H_1

→ If the calculated χ^2 is less than we accept H_0

Data Warehouse

- OLTP
- OLAP Systems

OLTP:- we are using in day to day operation.

OLAP:- we put historical data after integration/transformation and various other operations.

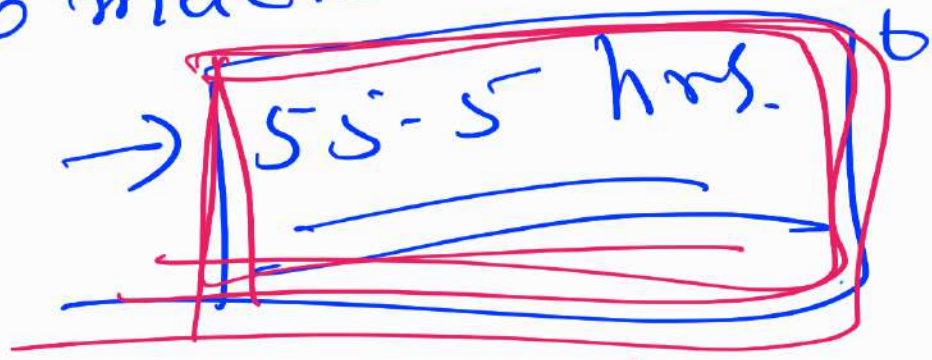
→ On-line analytical processing

→ 50 Transactions
↳ ≈ 1 sec

1 Tx. → $\frac{1}{50}$

1 Crore → $\frac{1}{50} \times 10000000$

Ifore much hours ?



OLAP ⇒ Analysis | Decision making

↳ Managers | Owners or other Policy makers

Total Revenue Generated in Patna by Airtel or other

→ Weekly

W1 [ ] Rev

W2 [ ] Rev2

Data warehouse :-

→ to systematically organize.
→ understand
→ use their data to make strategic decisions

William H. Inman :-

→ Subject Oriented
→ Integrated
→ Time-variant
→ Non-volatile.

in support of

management decision making process

Subject Oriented :-

↳ Centered around major subjects ex :- sales, suppliers

Integrated :-

↳ Integrating data from multiple sources. These ~~sour~~ sources may be homogeneous or hetrogeneous.

Time-Variants

↳ data is stored from the historical perspective

→ Every key structure in DW contain either implicitly

or explicitly an element of time.

Nonvolatile :-

→ we perform two operation

→ initial Loading of data

→ Access of data

Major Difference b/w OLTP & OLAP

OLTP → day to day operation

OLAP → used by decision makers

Data Contents :-

OLTP → Current data, detailed

OLAP → large amount of historical, aggregated

data.

Database Design

OLTP → ER models

OLAP → star or snowflake.
fact-constellation

Access Platform

OLTP → short, atomic Transactions

OLAP → read only operations

Data Cubes :-

→ Multidimensional Data models.

↓

Data Cubes

A data cube allows data to be modeled and viewed in multiple dimensions.

→ dimensions

→ facts

→ Entities wrt which an organization wants to keep records

For Sales subject

dim → time, items, branch, location etc.

→ Each dim^n has table associated with it

↓

dimension table.

→ The subject around

which multi dim^n data model is organized is represented by FACT table.

Ex:- 2D - view of sales Data

dim → time & item.

Sales from branches in Patner

location = "Patna"

time (Quarter)	item (type) Home Enter	Computer	Laptop	Mobile
Q1	609	961	106	-
Q2	514	102	-	-
Q3	623	103	107	-

Q3 —

Q4 691 104 — —

→ Can organize the data in 3D

location = "Patna" | location = "Hyderabad"

items | ITE | Com | Loc | ...

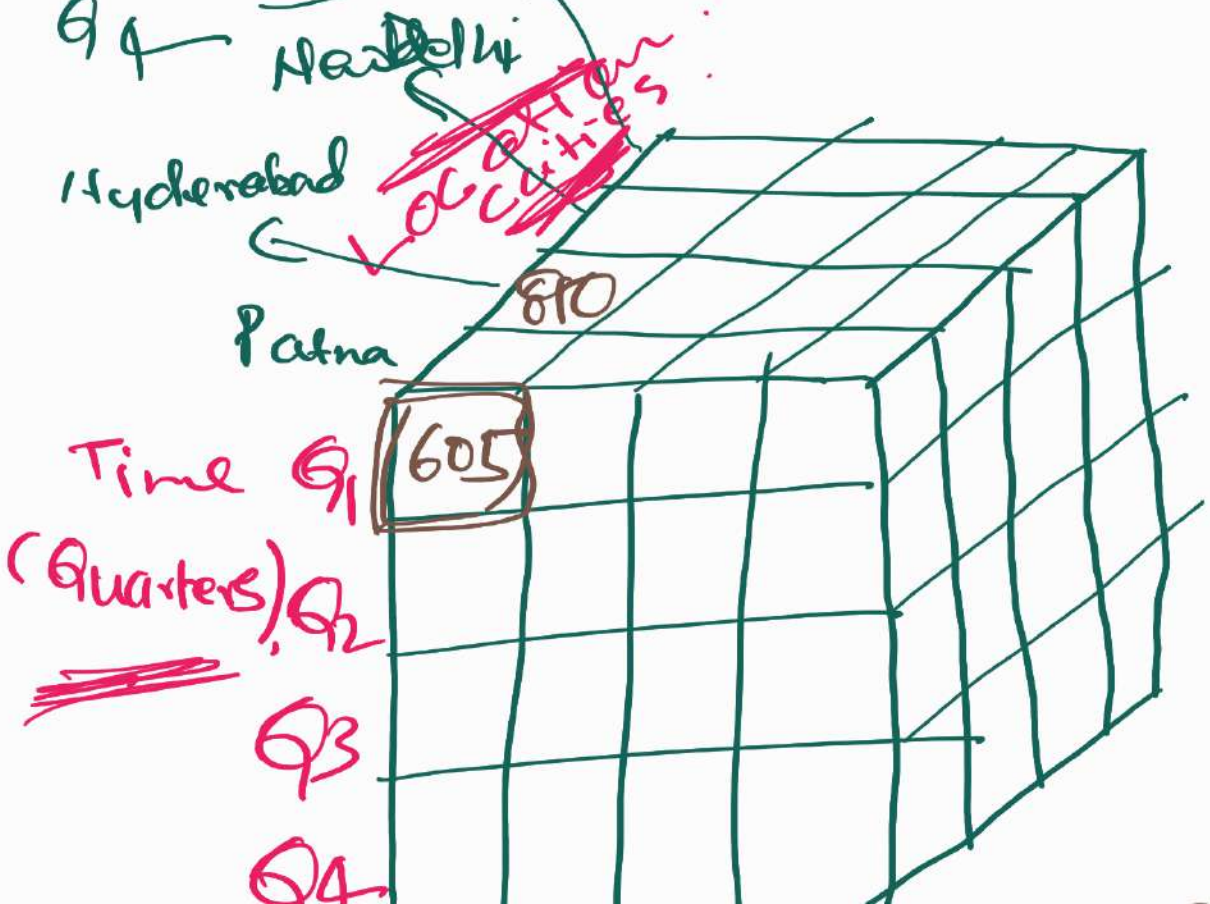
Q1 — — — —

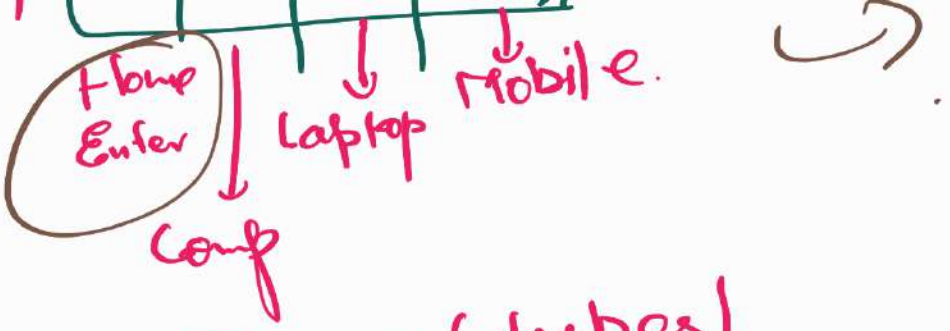
Q2 — — — —

Q3 — — — —

Q4 — — — —

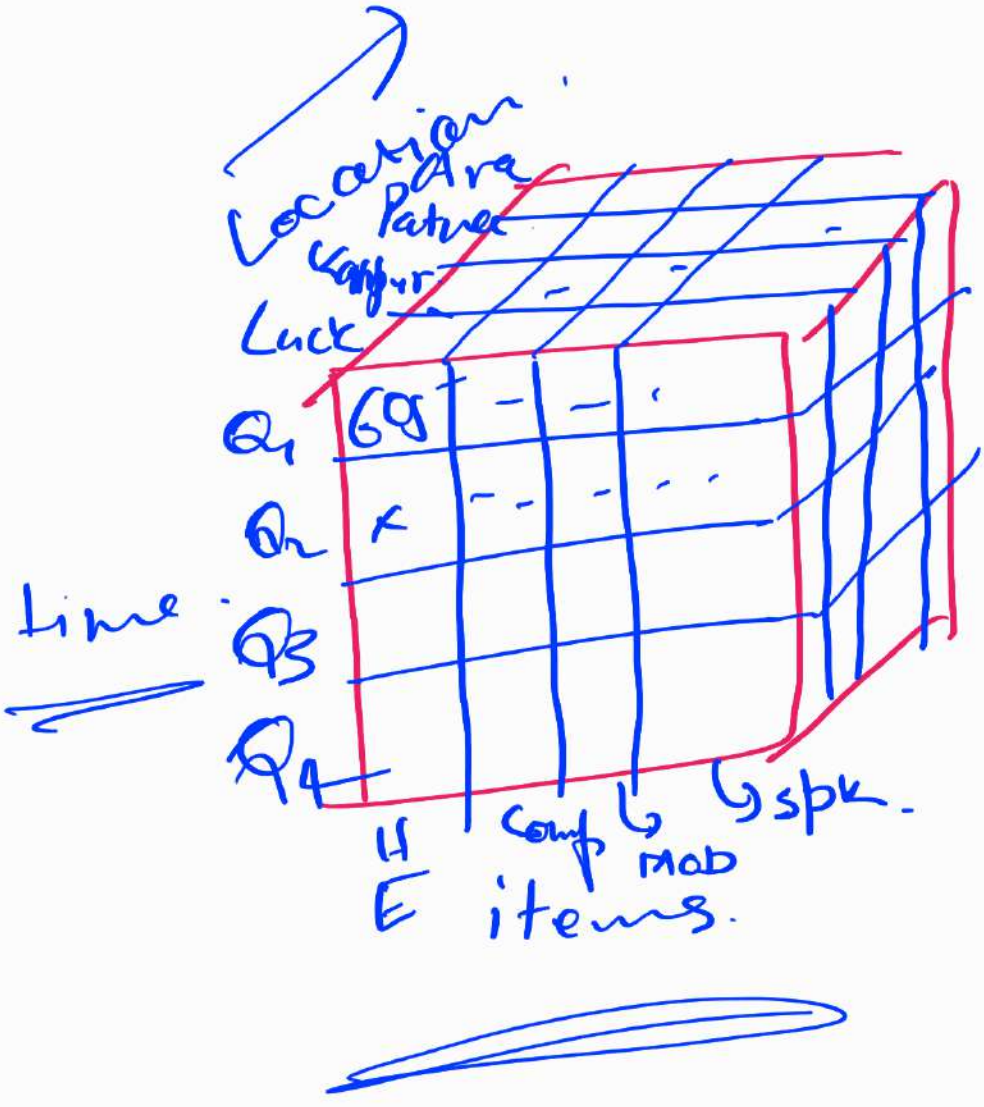
Add more cities.





Items (types).

Data Cube :-

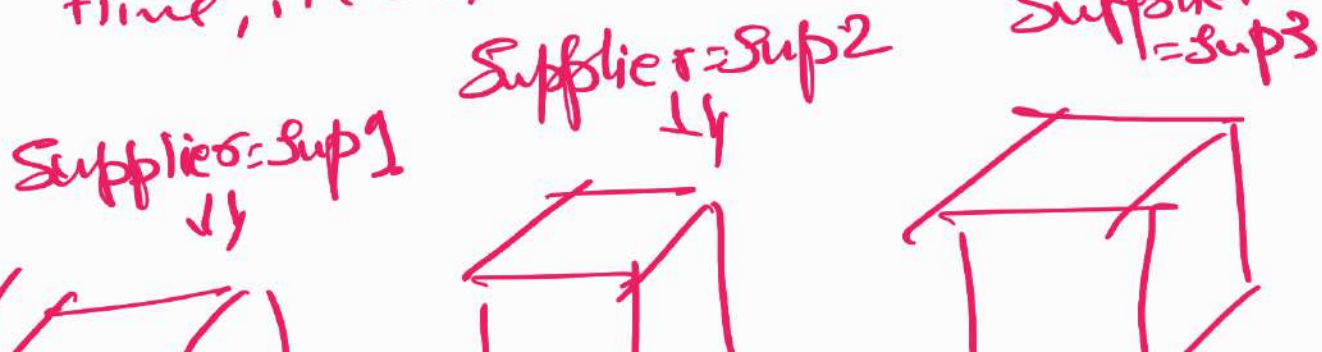


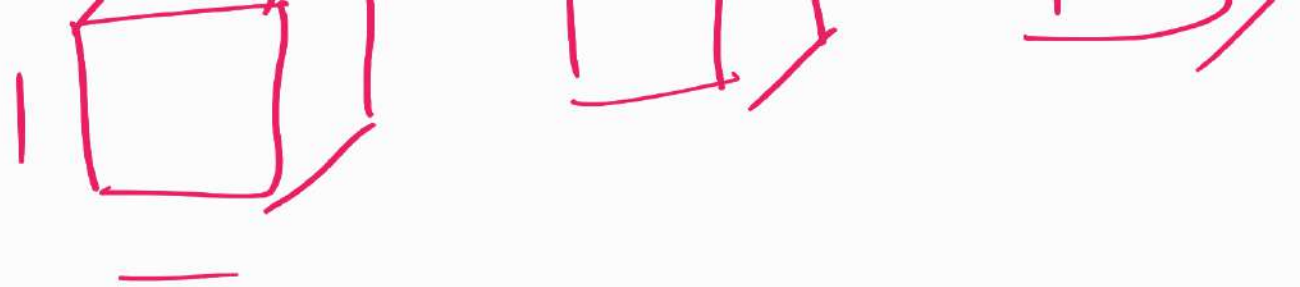
Four Dimn

→ As a series of 3D cubes

4th dim Supplier.

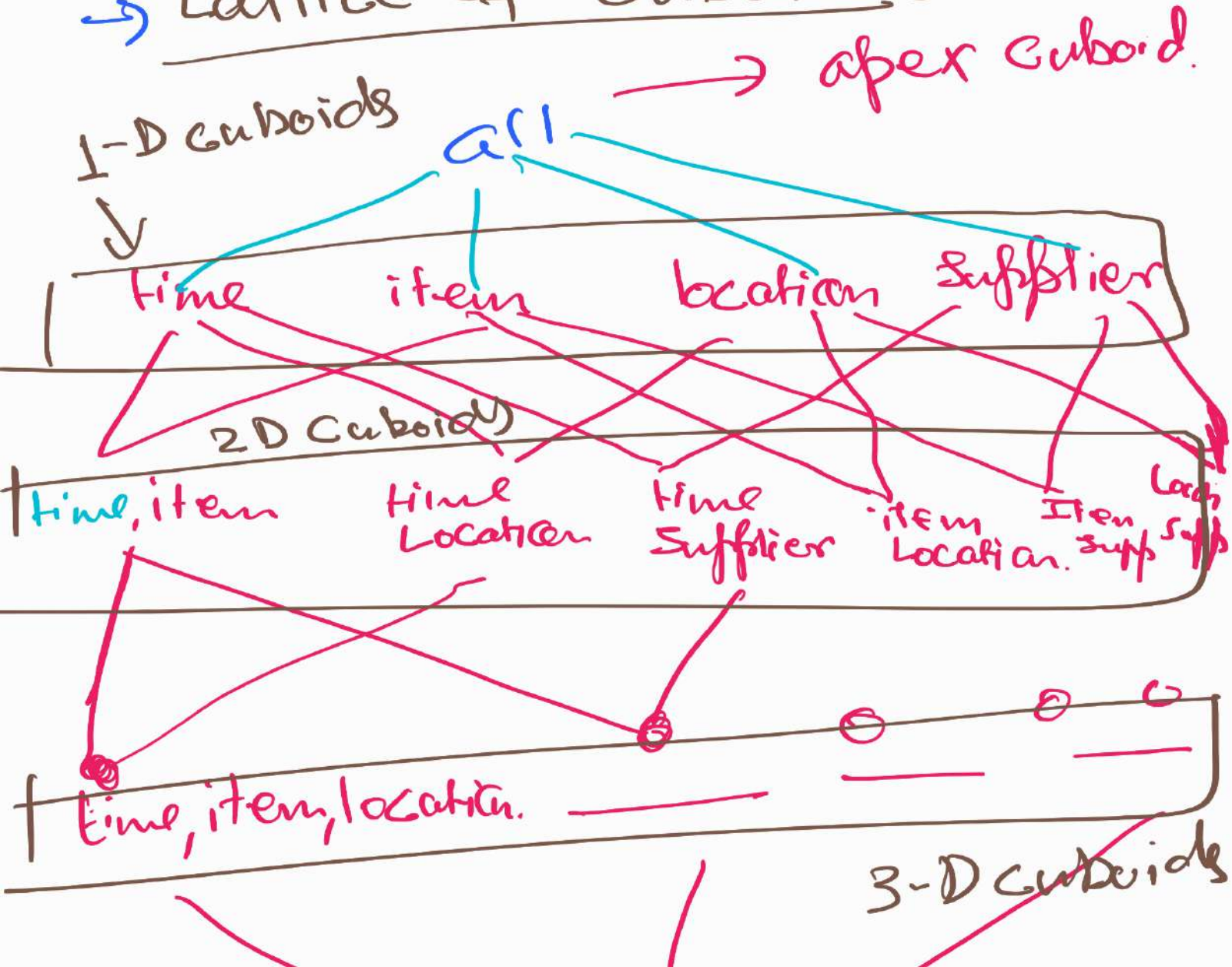
time, item, Location.





* If we have n-D data then we can represent it as a series of (n-1)-D cubes

→ Lattice of Cuboids :-

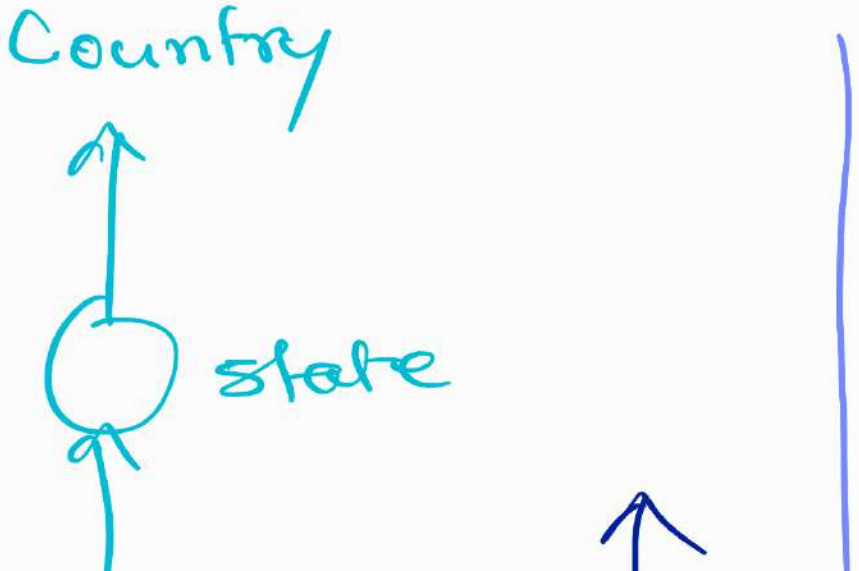


→ time, item, location.
Base Cuboid. supplier.

Lattice of cuboids

Concept Hierarchies

→ It defines a sequence of mapping from a set of low level concepts to higher level or more general concepts



Schemas for Multidimensional Databases

→ Star schema
→ Snowflake schema
→ Fact constellation schema.

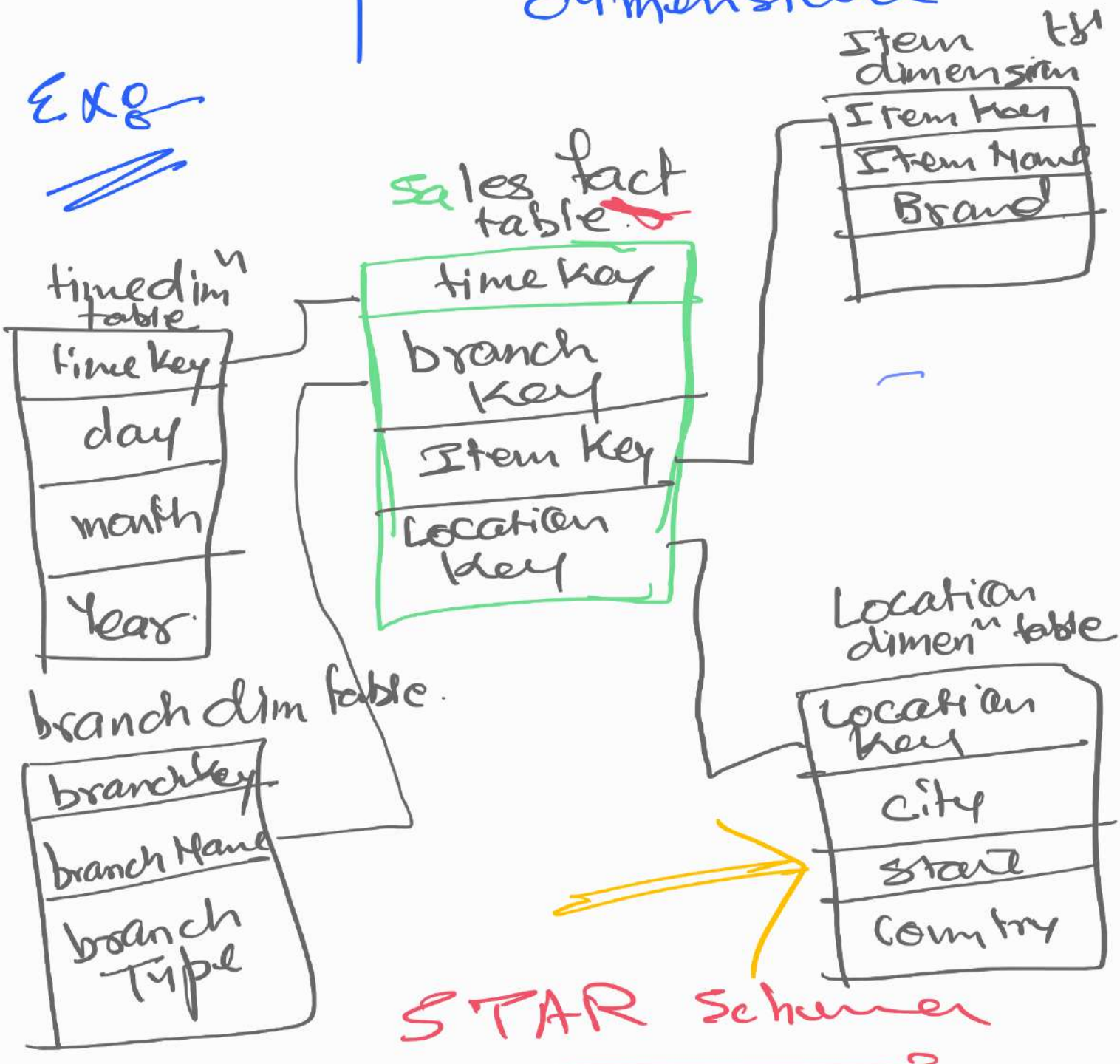
Star schema :-

→ It has large central table (fact table) ~~cont~~ containing the bulk of data with no redundancy

→ set of smaller attendant tables (dimension tables)

→ one for each dimension

Eg



STAR Schemes

for sales data.

Snowflake schemes -

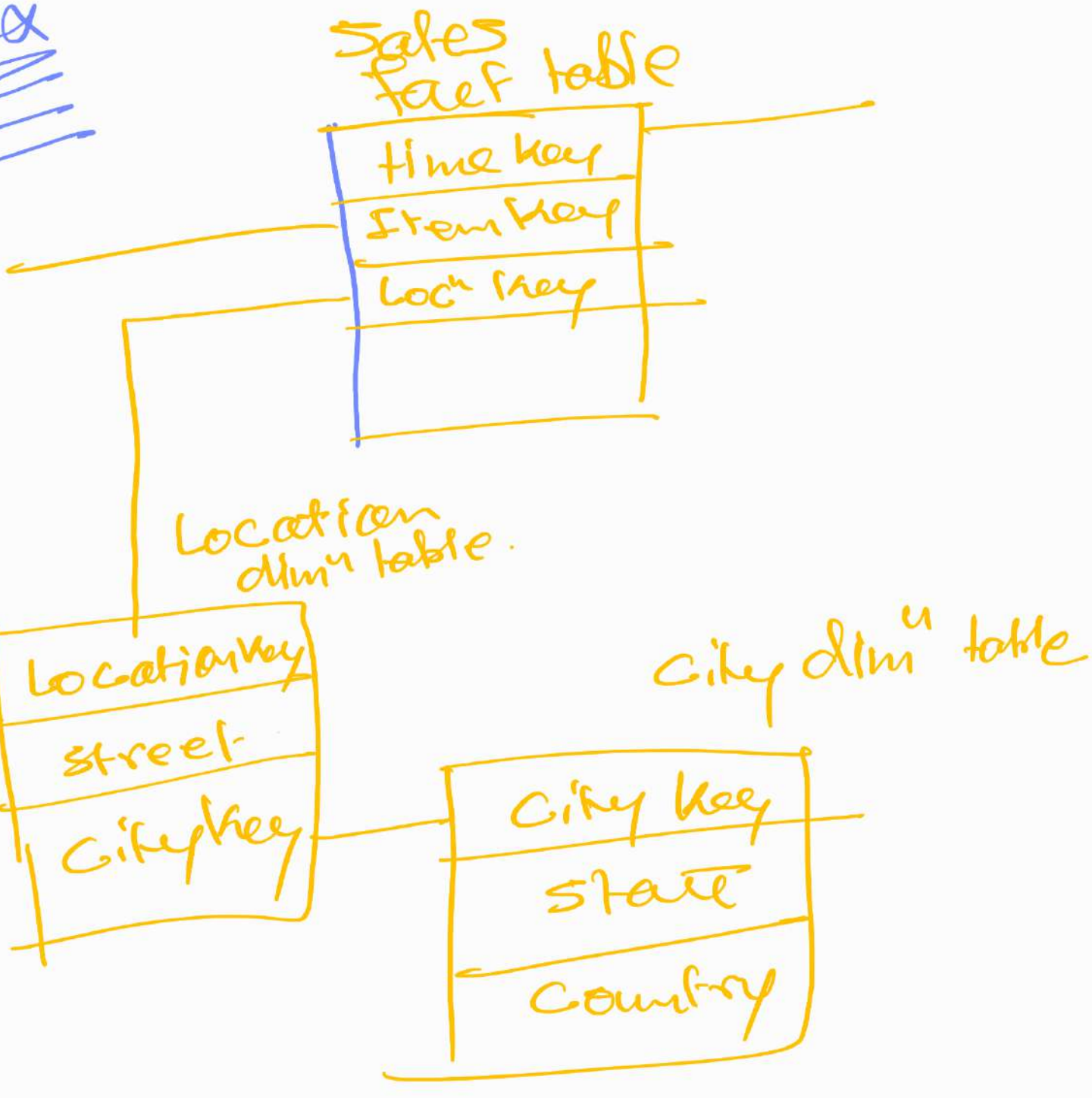
* Variant of STAR.

dimension table are

normalized.

→ Further splitting of data.

Ex



Fact Constellation

→ Galaxy schema

→ Sophisticated / complex aggⁿ

→ Has multiple fact tables that

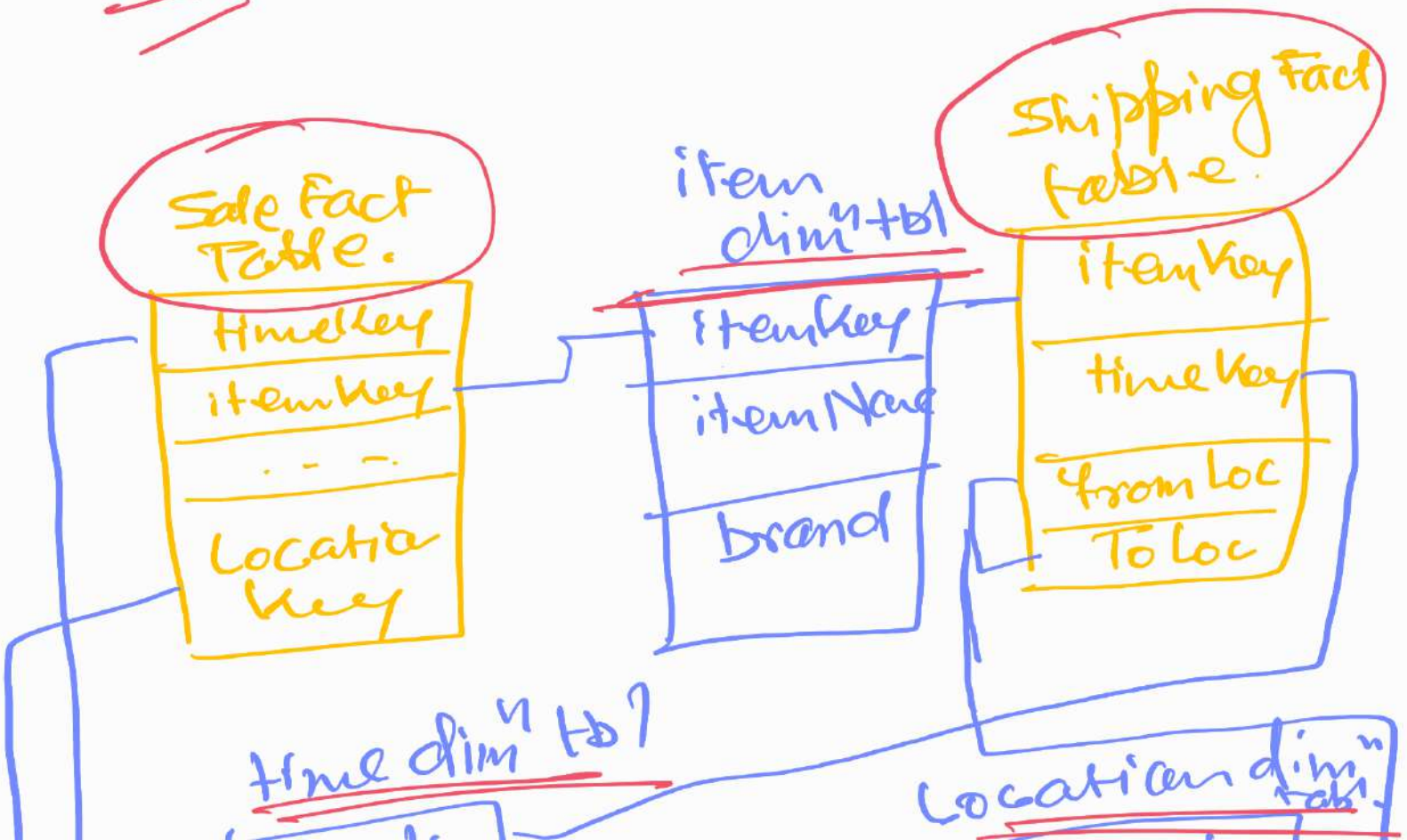
↓

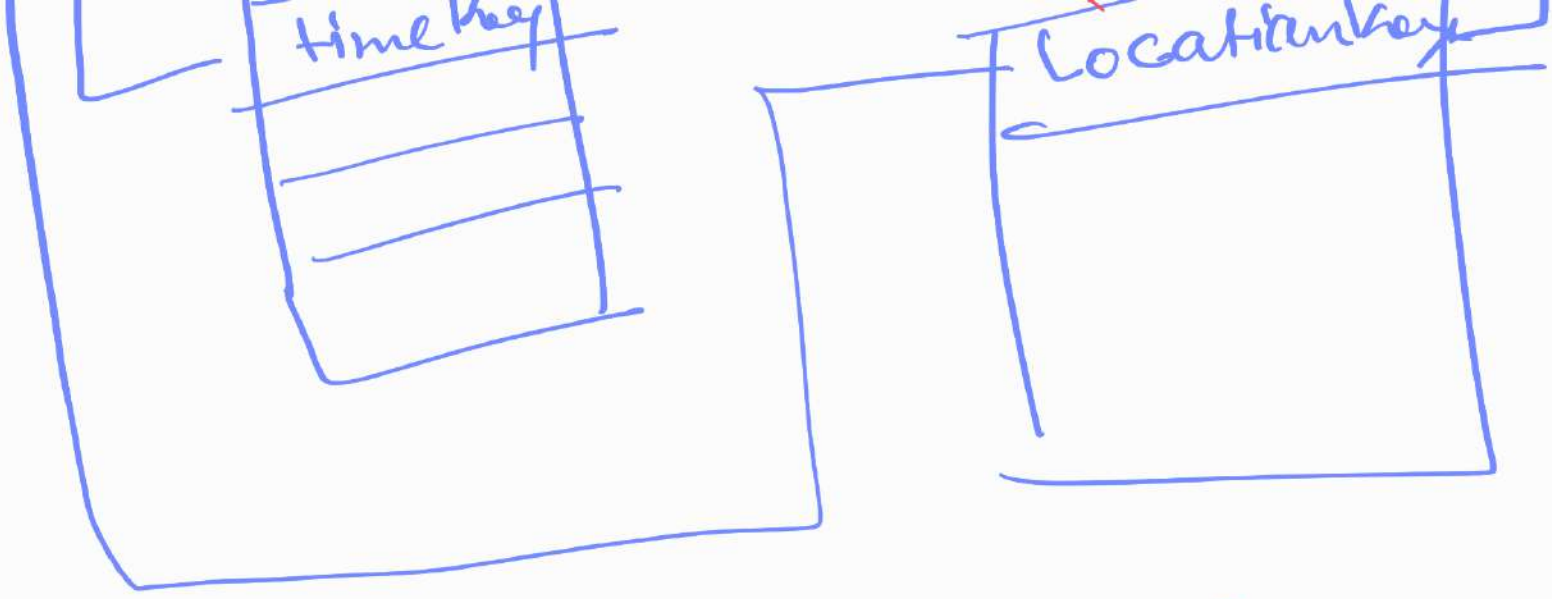
share

↓

dimension tables

Ex:





Fact constellation schema for sales & shipping

OLAP operations & —

→ Roll-up & —

→ performs aggregation

↓

data cube.

↓

climbing up the concept hierarchy

or dim^n reduction.

→ Drill Down :-

→ Reverse of roll-up

→ navigates from

↓

less detail

↓

more detail

→ moving down the concept hierarchy

→ Introduce additional dim^n

Slice & dice :-

Slice &

→ performs selection on one dim^{n}

Dice → defines subcube by performing selection on two or more dim^{n}

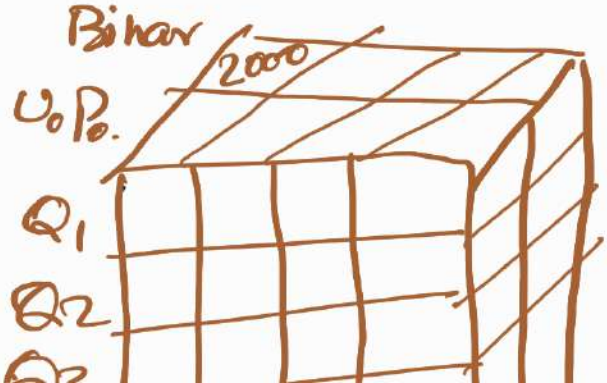
Pivot (rotate) & —

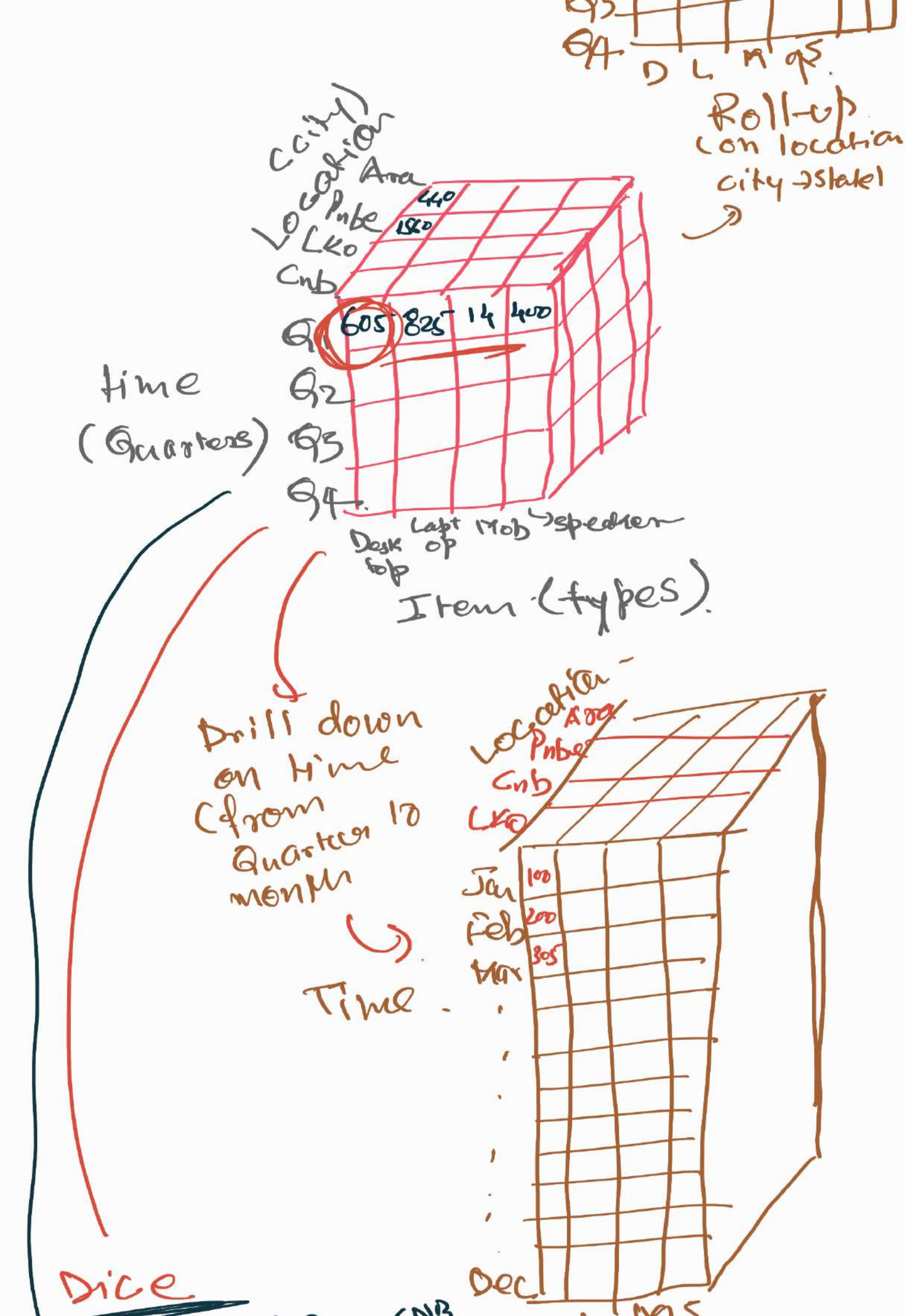
→ It rotates the data axis in view

↓

alternative presentation.

Example & —

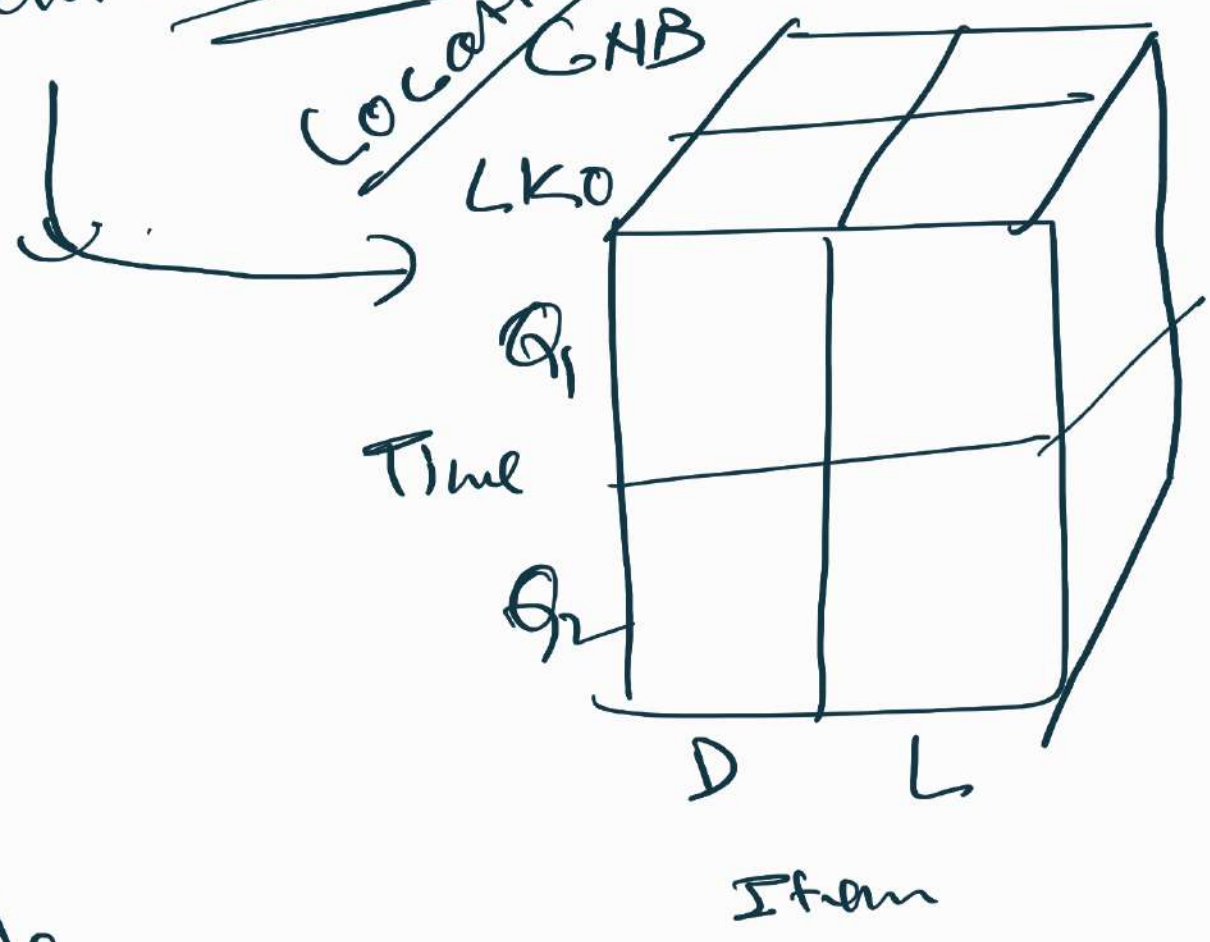



Roll-up
(on location
city → state)
City
Location
Area
Phone
LKO
Cnb
440
1820
605
825
14
400
Q1
Q2
Q3
Q4
time
(Quarters)
Desk top
Laptop
Mob
Speaker
Item (types)
Drill down
on time
(from
Quarters to
months
Time
Jan
Feb
Mar
100
200
305
Dec
Dice

Location = LKO or CHB

Time = Q_1 or Q_2

Item = D or L

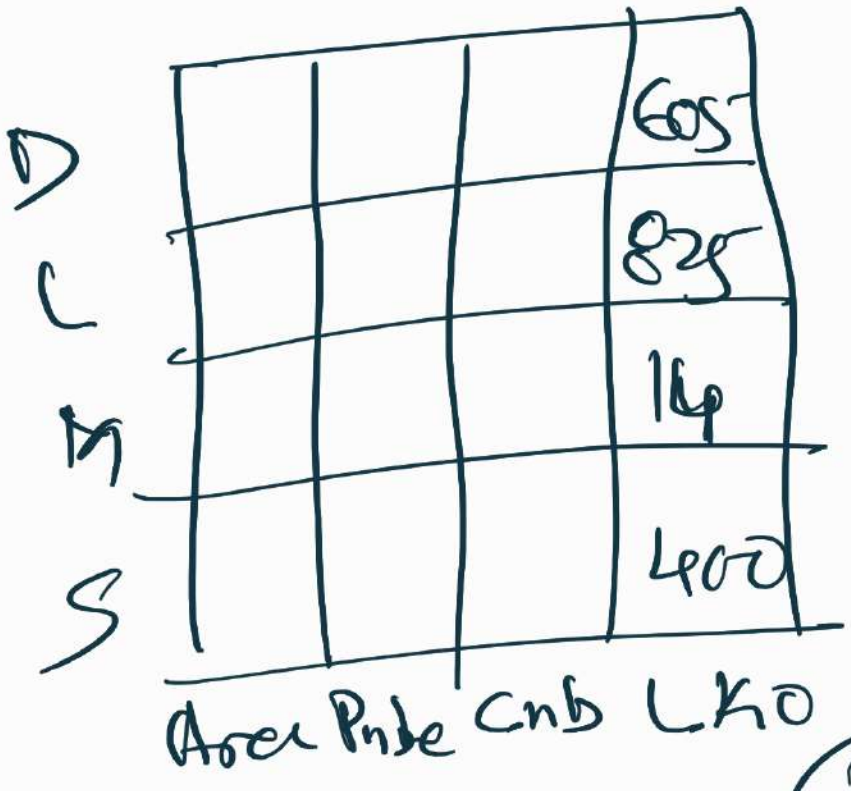


Slice

for Time = "Q1"

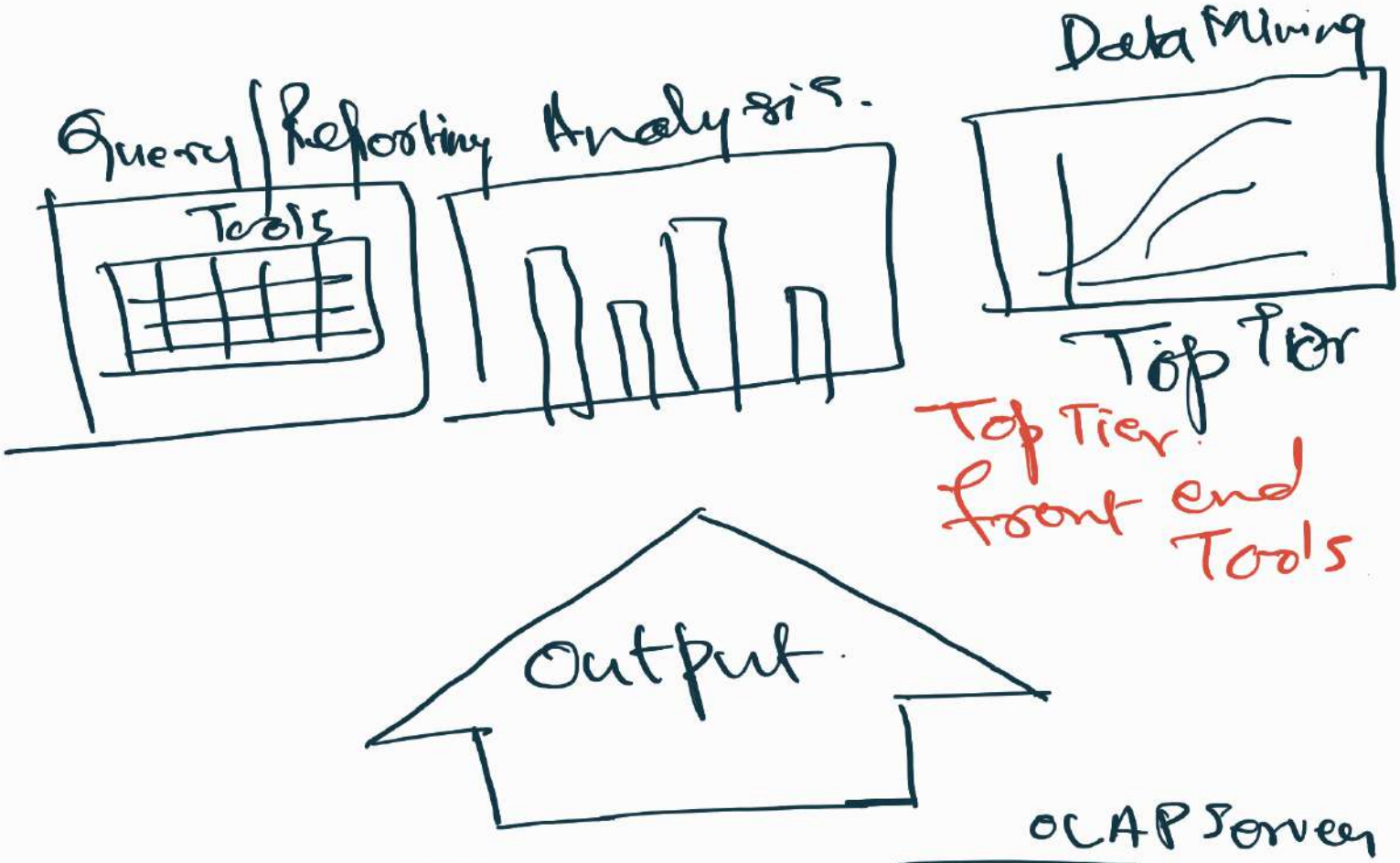
	D	L	TV	S.
Area				
Probe				
Car				
CKO	605	825	14	479

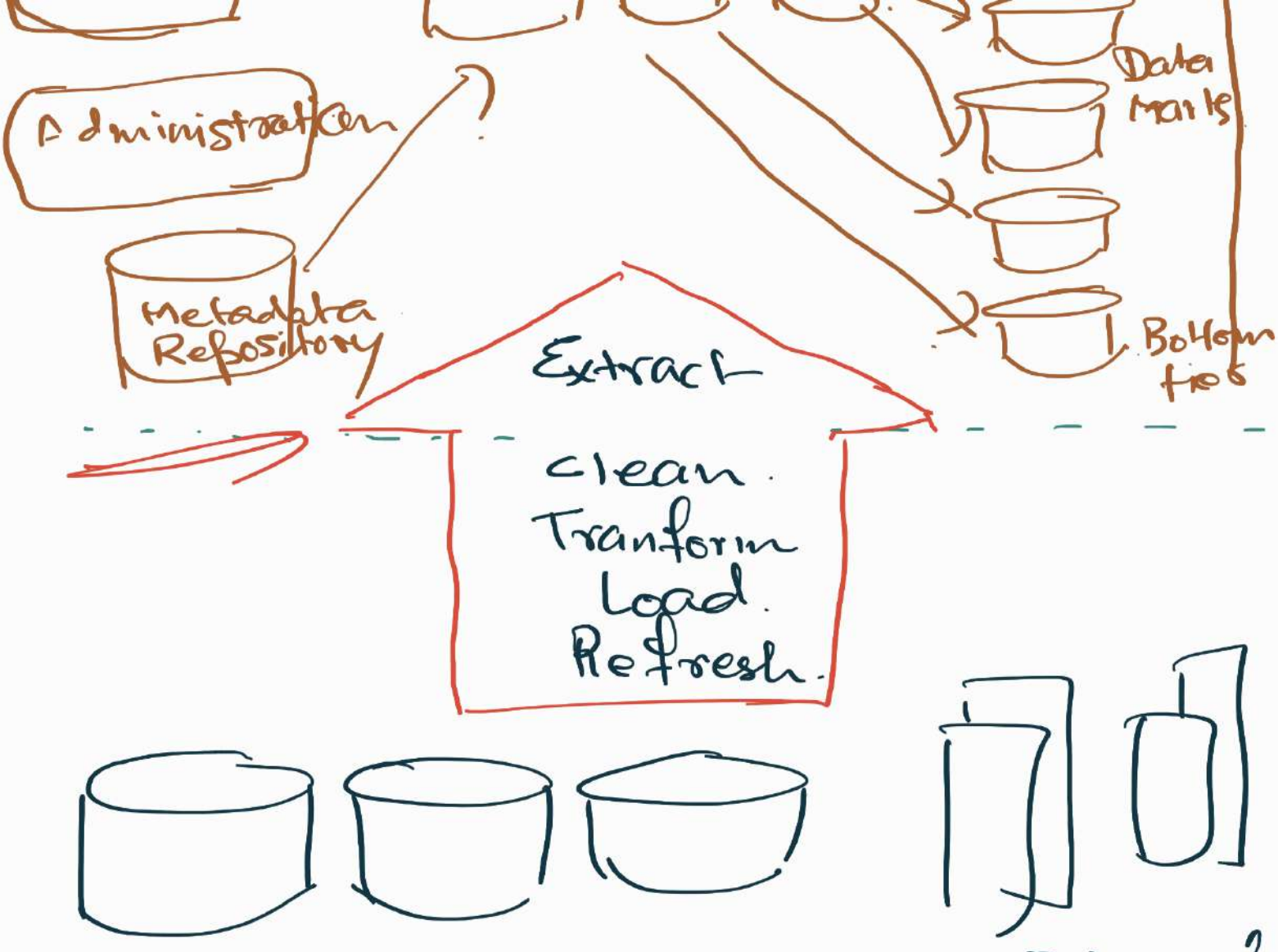
Pivot – (Rotate the axes to change the view)



→ Han & Kamber ① ②

Three-Tier Data Warehouse Archi°-





Three Tier data warehouse Archi°

Middle Tier° —

↓

If is an OLAP Server

→ ROLAP
→ MOLAP.

ROLAP (Relational OLAP)

→ extended form of relational DBMS

MOLAP :- (Multidimensional OLAP)

* special purpose server that directly implements multidim^{n} data & operations

Top Tier

↳ front end client

layer.

→ Query / reporting / Analysis tools

Enterprise Data warehouse

→ all information.
→ about various subjects.
→ entire organization.
→ cross functional in scope
→ Detailed & summarized data.
→ Terabytes

Data mart —

→ subset of corporate wide data.

→ specific group of users.

- → Independent
- → Dependent

Independent Data —

→ Data directly from the External information providers / other operational System.

Dependent Data —

→ data from the Enterprise data warehouse

Frequent Pattern

→ It changes in

⇒ dataset frequently

Ex milk & bread.

⇒ Subsequence

* sequence that occurs frequently in the dataset-

↳ frequent sequential Pattern

PC ⇒ Printer ⇒ Mouse Pad.

Market Basket Analysis :-

⇒ Ex. frequent itemset mining

→ Analysis customer buying habits

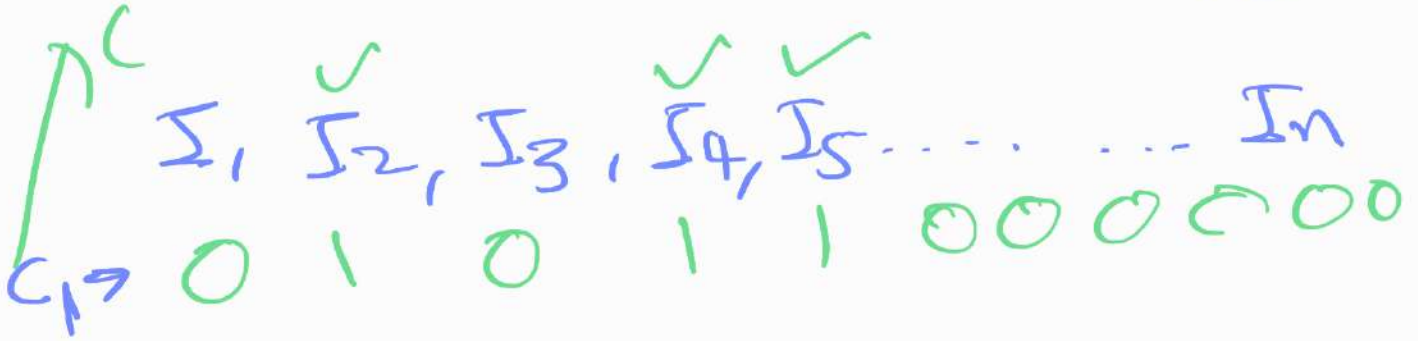
→ Association among diff items that he/she places in their shopping cart

→ Retailers to develop marketing strategies

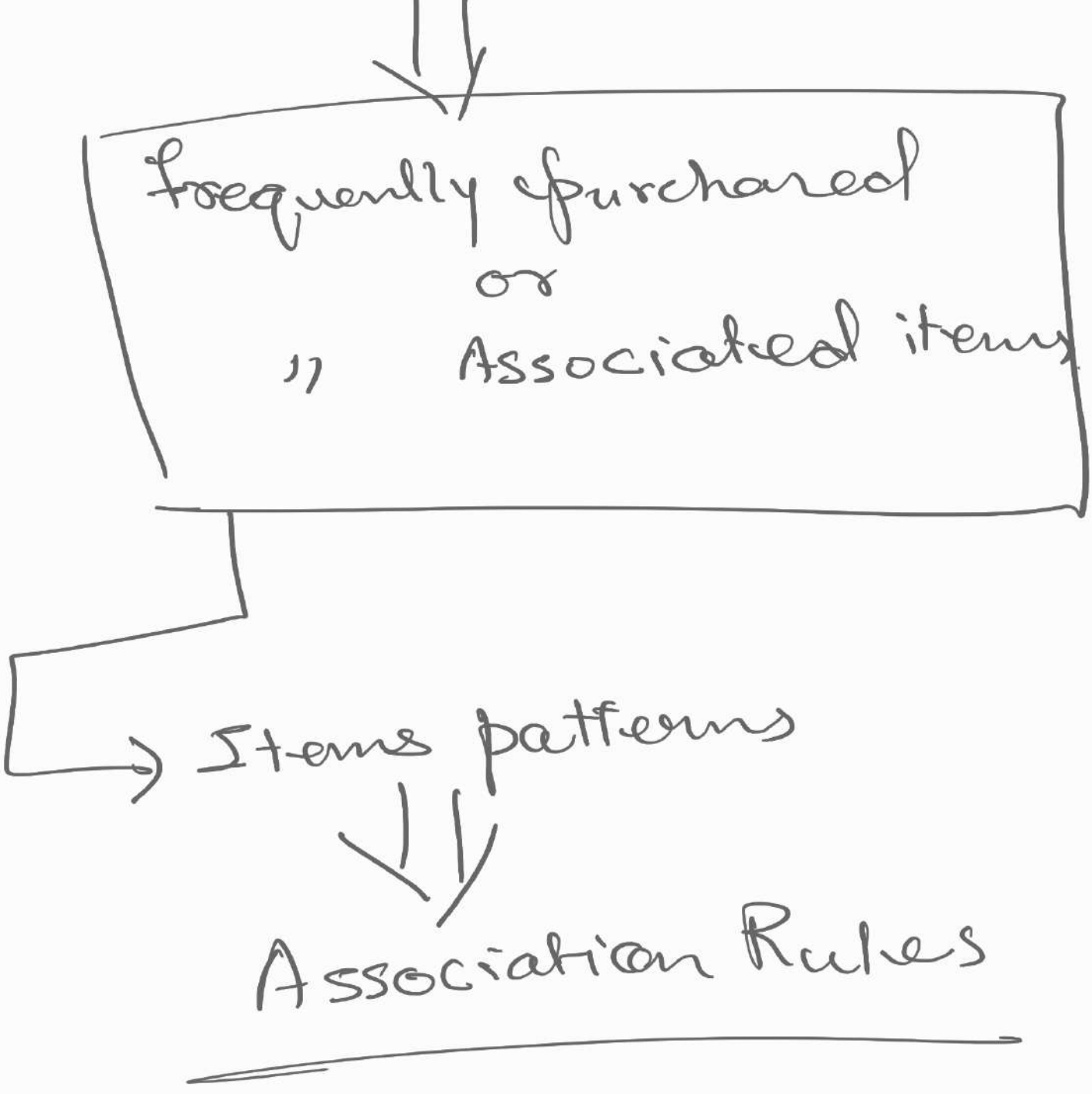
Association Rule :-

Each item → Boolean variable. (0/1)

Customer Basket → Boolean vector



* Boolean vector can then be analyzed for



Ex:- Computers, antivirus s/w

computer → Antivirus s/w

[support = 2% Confidence = 60%]

* → measuring the

Association rule interestingness

2% support :-

2% of all the transaction under analysis show that comp & antivirus s/w are purchased together.

60% confidence :-

→ 60% of the customers who purchased of computers also bought the software.

Antivirus

→ minimum support — threshold

→ minimum confidence threshold


A → B